

A new fast algorithm for reproducing complex networks with community structure

Mateusz Kowalczyk, Piotr Fronczak, Agata Fronczak

Faculty of Physics, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

Abstract

In this paper we introduce a new algorithm allowing for generation of networks with heterogeneity of both node degrees and community sizes. The quality and efficiency of the algorithm is analyzed and compared to the other, so far the most popular algorithm which was proposed by Lancichinetti et al. We discuss the advantages and shortcomings of both algorithms indicating the areas of their potential application.

Keywords: complex networks, community structure, algorithms

1. Introduction

The community structure is considered to be, next to the small-world effect and scale-free degree distribution, one of the most important topological properties of real networks. By the community (also called cluster, module, or block) in a network we understand a group of nodes more densely
 5 connected to each other than to nodes outside the group. For example, in social networks, communities correspond to groups of people sharing the same interests [1], in Internet, they consist of the sets of web pages on the same topic [2], while in cellular and metabolic networks, communities are
 10 functional modules of interacting proteins [3].

In the science of complex networks, community detection has become one of the most dominant research topics over the last decade. As a consequence,

Email address: `fronczak@if.pw.edu.pl` (Mateusz Kowalczyk, Piotr Fronczak, Agata Fronczak)

a large number of algorithms have been proposed for the analysis of community structure in network [4, 5, 6, 7]. To evaluate these algorithms effectively, synthetic networks with a well-defined community structure (benchmarks) had to be proposed. The advantage of such models is that, unlike in real networks, one can easily vary the model parameters and compare the recovered community structure with the predefined one.

One of the first models of networks with community structure, with a long tradition of study in the social sciences and computer science [8, 9, 10, 11, 12, 13], is the so-called blockmodel. In its classical version [8], each of N nodes is assigned to one of K blocks (communities) of equal size, and undirected edges are independently drawn between pairs of nodes with probabilities that are a function only of the group membership of the nodes. Unfortunately, the Poisson-like degree distribution makes this model unsuitable for the further analysis, since most of real networks exhibit power laws in their degree distributions.

Lancichinetti et al. [14] proposed an efficient numerical construction procedure for benchmark graphs that is free of this defect. The method accounts for the heterogeneity in the distributions of node degrees as well as community sizes. Its efficiency has been tested and proved in typical cases, however, further in this paper we show that in a certain range of parameters efficiency of the algorithm drops significantly. Moreover, the complexity of the proposed procedure does not allow for the analytic tractability.

In opposite to Lancichinetti's procedure, Fronczak et al. [15] provided an exponential random graph formulation [16, 17, 18, 19, 20] for blockmodel that is solvable for its parameter values in closed forms. Two kinds of the network structural Hamiltonians have been considered: the first one corresponding to the classical blockmodel, and the second one corresponding to its degree-corrected version. In both cases, a number of analytical predictions about various network properties was given. In particular, it was shown that in the degree-corrected blockmodel, node degrees display an interesting scaling property, that is similar to the scaling feature of the node degrees in fractal (self-similar) real-world networks. Unfortunately, the method is computationally inefficient since it is based on Markov chain Monte Carlo algorithm.

In this contribution we propose a simple, analytically tractable, and fast algorithm for generation of networks with community structure and heterogeneity of both node degrees and community sizes. The method allows to generate, in a reasonable time, networks that are orders of magnitude larger

than those generated by the previous approaches. It also allows for closed-form parameter solutions.

In outline, the paper is as follows. First, we introduce a new method (KA; the meaning of this abbreviation is "Kowalczyk's et al. algorithm") for generating clustered networks and derive their main properties. Next, we review Lancichinetti's algorithm (LA). We describe its sub-procedures and their time complexity. This allows us to point the range of parameters for which the algorithm efficiency drastically drops down. Finally, we discuss all the major pros and cons of the both approaches. In the appendix, we provide detailed listings of the both algorithms.

2. Derivation of the new algorithm

In this section, we present a simple algorithm to generate networks with community structure, which, despite its simplicity, has not been considered, at least to our knowledge, in previous studies. The algorithm is an extension of the model for generating uncorrelated networks with a given sequence of expected degrees $\{\langle k_1 \rangle, \langle k_2 \rangle, \dots, \langle k_N \rangle\}$ (see eg. Eq. (15) in [21] and Eq. (48) in [17]). In such a prototype network, there is at most one link between any pair of nodes, and there are no self-loops connecting nodes to themselves. If a_{ij} is an entry of the adjacency matrix underlying the network, and $a_{ij} \in \{0, 1\}$, where $a_{ij} = a_{ji}$ and $a_{ii} = 0$, then the expected value of the entry, $\langle a_{ij} \rangle$, can be expressed in terms of the probability, p_{ij} , that the vertices i and j are connected, namely

$$\langle a_{ij} \rangle = 1 \cdot p_{ij} + 0 \cdot (1 - p_{ij}) = p_{ij}. \quad (1)$$

Simultaneously, given the expected node degrees, the average number of connections, which obviously can not be greater than one, may be estimated as the expected number of successes in $\langle k_i \rangle$ attempts of i to connect to j , where the probability of success for one trial is $\langle k_j \rangle (\sum_{j \neq i} \langle k_j \rangle)^{-1}$, i.e.

$$\langle a_{ij} \rangle = \langle k_i \rangle \frac{\langle k_j \rangle}{\sum_{j=1}^N \langle k_j \rangle - \langle k_i \rangle} \simeq \frac{\langle k_i \rangle \langle k_j \rangle}{\langle k \rangle N}. \quad (2)$$

By comparing Eqs. (1) and (2), one gets a simple expression for the probability of a connection:

$$p_{ij} = \frac{\langle k_i \rangle \langle k_j \rangle}{\langle k \rangle N}. \quad (3)$$

In analogy to the above derivation, in networks with community structure, one can write similar relations for the probabilities p_{ij}^{int} and p_{ij}^{ext} , that there is an internal or external connection between two nodes, i and j , belonging to the same or to different communities. If it is not clear, let us explain that internal connections are those that are between nodes belonging to the same community. Accordingly, the external connections are those that are between nodes belonging to different clusters.

Thus, let $\langle k_{i,r}^{int} \rangle$ represent the expected internal degree of a node i belonging to the r -th community. Correspondingly, let $\langle k_{i,r}^{ext} \rangle$ be the expected number of its external connections. Then:

$$p_{ij}^{int} = \langle k_{i,r}^{int} \rangle \frac{\langle k_{j,r}^{int} \rangle}{\sum_{j=1}^{c_r} \langle k_{j,r}^{int} \rangle - \langle k_{i,r}^{int} \rangle} \simeq \frac{\langle k_{i,r}^{int} \rangle \langle k_{j,r}^{int} \rangle}{2 \langle E_r^{int} \rangle}, \quad (4)$$

and

$$p_{ij}^{ext} = \langle k_{i,r}^{ext} \rangle \frac{\langle k_{j,s}^{ext} \rangle}{\sum_{s \neq r} \sum_{j=1}^{c_s} \langle k_{j,s}^{ext} \rangle} \simeq \frac{\langle k_{i,r}^{ext} \rangle \langle k_{j,s}^{ext} \rangle}{2 \langle E^{ext} \rangle}, \quad (5)$$

where c_r is the size of the r -th cluster, $\langle E_r^{int} \rangle$ is the expected number of internal links within r , and $\langle E^{ext} \rangle$ is the number of external links in the whole network.

Now, let the mixing parameter, μ , describe a share of links which connect each node with nodes belonging to other clusters, i.e.

$$\langle k_{i,r}^{ext} \rangle = \mu \langle k_{i,r} \rangle, \quad (6)$$

and

$$\langle k_{i,r}^{int} \rangle = (1 - \mu) \langle k_{i,r} \rangle, \quad (7)$$

where

$$\langle k_{i,r} \rangle = \langle k_{i,r}^{int} \rangle + \langle k_{i,r}^{ext} \rangle, \quad (8)$$

is the expected total degree of the node i , which belongs to the cluster r . Using Eqs. (6)-(8) the connection probabilities, Eqs. (4) and (5), can be rewritten as follows:

$$p_{ij}^{int} = \frac{(1-\mu) \langle k_{i,r} \rangle (1-\mu) \langle k_{j,r} \rangle}{(1-\mu) \sum_{j=1}^{c_r} \langle k_{j,r} \rangle} = \frac{(1-\mu) \langle k_{i,r} \rangle \langle k_{j,r} \rangle}{\langle k \rangle c_r}, \quad (9)$$

and

$$p_{ij}^{ext} = \frac{\mu \langle k_{i,r} \rangle \mu \langle k_{j,r} \rangle}{\mu \sum_{j=1}^N \langle k_{j,r} \rangle} = \frac{\mu \langle k_{i,r} \rangle \langle k_{j,r} \rangle}{\langle k \rangle N}, \quad (10)$$

where it has been assumed that the average degree of the nodes within each community is the same as the average degree averaged across the whole network, i.e.

$$\langle k \rangle = \frac{1}{c_r} \sum_{j=1}^{c_r} \langle k_{j,r} \rangle = \frac{1}{N} \sum_{r=1}^n \sum_{j=1}^{c_r} \langle k_{j,r} \rangle, \quad (11)$$

where n is the number of clusters.

Having the probabilities p_{ij}^{int} and p_{ij}^{ext} derived, one can generate networks with community structure using the following algorithm:

- 75 1. For each node v draw an expected degree $\langle k_v \rangle$ from a power distribution $P_\gamma(k) \sim k^{-\gamma}$.
2. Generate n clusters with sizes c_r drawn from a power distribution $P_\beta(c) \sim c^{-\beta}$. Assign each created cluster to c_r consecutive nodes. The sum of all cluster sizes should not be smaller than the number N of nodes in the network.
- 80 3. For each pair of nodes (i, j) add a link with the probabilities given by Eqs. (9) or (10) depending on whether or not the two nodes share the same cluster.

The algorithm is listed in detail as the **Algorithm 1** in the Appendix.

- 85 Before we discuss the quality and efficiency of the presented algorithm we would like to restate the algorithm LA, which was provided by Lancichinetti et al. in Ref. [14]. Nowadays, LA is one of the most frequently cited method for generating clustered network in the literature. Having the both algorithms presented we will be able to compare their advantages and shortcomings which will give one a reference point to decide by himself which algorithm
- 90 would fit better to the specified needs.

3. The algorithm introduced by Lancichinetti et al.

- Here we restate the LA algorithm and discuss some of the implementation issues that significantly impact its performance. We follow the same notation as in the original article [14]. In particular, we assume that the node degrees are drawn from a power-law distribution with the exponent γ , the community sizes are drawn from a power-law distribution with the exponent β , the number of nodes is N , the minimal, average, and the maximal degree are: k_{min} , $\langle k \rangle$, and k_{max} , respectively. Furthermore, the mixing parameter μ ,
- 95

100 as in previous section, describes the share of links that connect each node with nodes belonging to other communities.

The algorithm comprises of several steps that are listed as **Algorithm 2** in the Appendix.

1. For each node v draw a degree k_v from the power-law distribution
105 $P_\gamma(k) \sim k^{-\gamma}$.
2. Assign an expected internal degree $\langle k_v^{int} \rangle$ to each node v according to relation $\langle k_v^{int} \rangle = (1 - \mu)k_v$, cf. Eq. (7). Please note, that, in opposite to the total node degrees, the internal degrees obtained in numerical simulations, k_v^{int} , may differ from the expected values, $\langle k_v^{int} \rangle$; the latter
110 can only be realized in average.
3. Create an initial network using the so-called configuration model [22]. In this model, at the beginning, exactly k_v "stubs" or half-edges emanate from each node v . Then, the network is constructed by choosing a uniformly random matching on these degree stubs. It is worth to note,
115 than the obtained networks can contain self-loops and multi-edges (i.e. they are multigraphs). These represent usually a tiny fraction of all edges, and one can just discard or collapse them, however for $\gamma < 3$ this operation can lead to the so-called structural correlations.
4. Generate empty clusters with capacities drawn from the power-law distribution $P_\beta(c) \sim c^{-\beta}$. The sum of all capacities should not be smaller than the number N of nodes in the network.
5. Assign nodes to clusters. Initially empty clusters are successively filled by the nodes under assumption that the internal degree of the inserted node can not exceed the cluster capacity. If the cluster is full (i.e. when
125 its size equals its capacity), then before inserting a new node, one of the nodes previously assigned to this cluster is removed. This step, as the most affecting the performance of the algorithm is described in detail as **Algorithm 3** in the Appendix.
6. Perform N/n steps (n is the number of clusters) of the optimization process that tries to minimize deviation between the actual internal degree, k_v^{int} , and the expected one, $\langle k_v^{int} \rangle$, namely

$$\sigma^2 = \sum_v (\langle k_v^{int} \rangle - k_v^{int})^2. \quad (12)$$

During each step the network configuration is updated via the link rewiring process, which preserves the degree of each node and affects
130 internal degrees only.

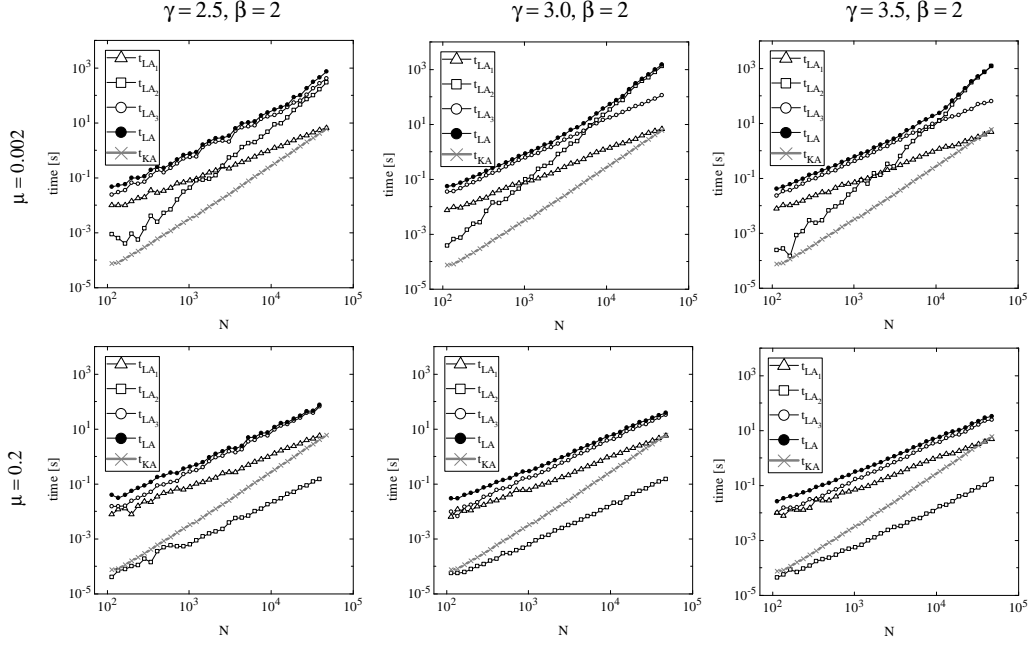


Figure 1: Comparison of the total execution time, t_{KA} , for the KA algorithm with the total execution time, t_{LA} , and the partial times t_{LA1} , t_{LA2} , and t_{LA3} corresponding to the specified sub-procedures of the LA algorithm. The figure presents data averaged over 10 realizations of networks with $\beta = 2$, $\langle k \rangle = 16$, $k_{max} = \sqrt{\langle k \rangle N}$, and k_{min} given by the normalization condition $k_{min} = \langle k \rangle (\gamma - 2) / (\gamma - 1)$, and different settings of the parameters γ and μ .

4. Comparative analysis of the two algorithms

The complexity $\mathcal{O}(N^2)$ of the KA algorithm is obvious due to the iteration over $\binom{N}{2}$ pairs of nodes which can be optionally connected. This complexity does not depend on any other parameter of the model. The execution time t_{KA} of this algorithm for different settings of the parameters γ , β , and μ is presented in Figs. 1 and 2.

The efficiency of the LA algorithm has been partially analyzed in [14]. The authors state therein that the procedure allows one to build fairly large networks (up to 10^5 - 10^6 nodes) in a reasonable time. Extracting data from Fig. 2 in Ref. [14], one can actually draw such a conclusion. However, as we will show later in this section, the time needed to build such large networks

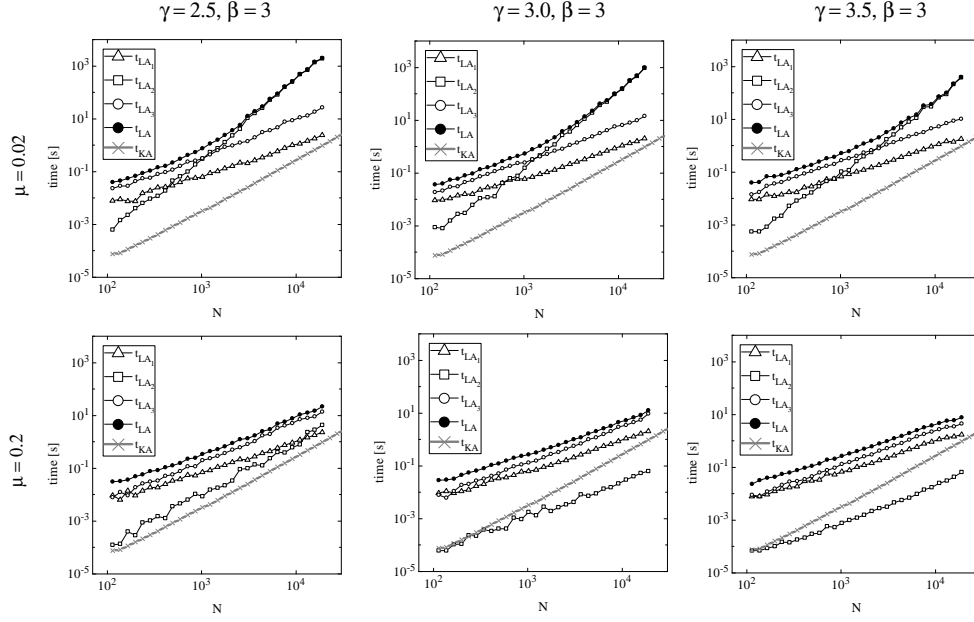


Figure 2: The execution time t_{LA} of the LA algorithm, times t_{LA_1} , t_{LA_2} , and t_{LA_3} of its sub-procedures, compared to the execution time t_{KA} of the KA algorithm for $\beta = 3$ and different sets of the parameters γ and μ . All the presented results are obtained similarly to those in Fig. 1.

may vary from 30 minutes to 20 days (on a 2.6 GHz Intel Core i5) depending on the choice of parameters γ , β , and μ .

145 To show this, we have analyzed execution times t_{LA_1} , t_{LA_2} , and t_{LA_3} , of the three main sub-procedures comprising the above algorithm (corresponding to the steps 3, 5, and 6 of the construction procedure, which is described in Sect. 3). We have omitted the analysis of other sub-procedures, since they have no visible impact on the total execution time t_{LA} .

150 Regarding the time t_{LA_1} needed to build the configuration model one can estimate its complexity as $\mathcal{O}(N)$. This scaling results from the number of "stubs" that have to be connected, which is twice a number of links $E = \langle k \rangle N$. The complexity $\mathcal{O}(N)$ of the time t_{LA_3} is simply due to the execution of N/n iterations in the sub-procedure 6. Both these predictions have been
155 confirmed experimentally for different sets of the parameters γ , β , and μ (see Fig. 1 and 2).

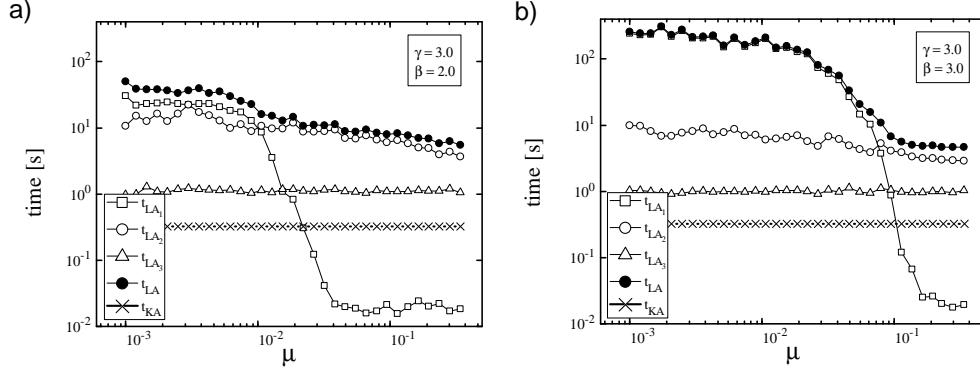


Figure 3: The execution time t_{LA} of the LA algorithm, times t_{LA_1} , t_{LA_2} , and t_{LA_3} of its specified sub-procedures as compared to the execution time t_{KA} of the KA algorithm for $N = 10000$ and different settings of the parameters γ and β . All the presented results are obtained similarly to those in Fig. 1.

The most interesting part of the LA algorithm takes place during the assignment of nodes to clusters. In the best case, when each node is assigned to its cluster without hindrance, the complexity of the time t_{LA_2} is simply linear with the system size, N . This usually happens when the clusters are large enough to include any node regardless of its expected internal degree. Such a situation can occur in two different ways. First, for sufficiently small expected internal degrees, i.e. for $\mu \rightarrow 1$, what corresponds to fuzzy communities, and second, when all cluster capacities are larger than k_{max} , what corresponds to the network consisted of only several large communities.

In the worst case, when all the cluster capacities are comparable with the expected internal degrees of nodes, the algorithm iterates $3N$ times trying to find an appropriate cluster for each node (what executes $3N * N$ times in total). If, after those $3N$ trials, there are still unassigned nodes, then the two smallest clusters are merged and the whole process repeats.

To see how does it work, let us shortly discuss the case of $\gamma \approx \beta$, i.e., when the both, node degrees and cluster capacities, are drawn from the same distribution. Then, the average cluster capacity $\langle c \rangle \approx \langle k \rangle$, and the merging process can repeat n times, where n is the number of clusters $n \approx N/\langle c \rangle \approx N/\langle k \rangle$. Taking all these iterations into account, one can estimate the complexity of the time t_{LA_2} as $\mathcal{O}(N^3)$. As one can see in Fig. 1 and 2, the time t_{LA_2} becomes a dominant factor for the whole processing time t_{LA} for

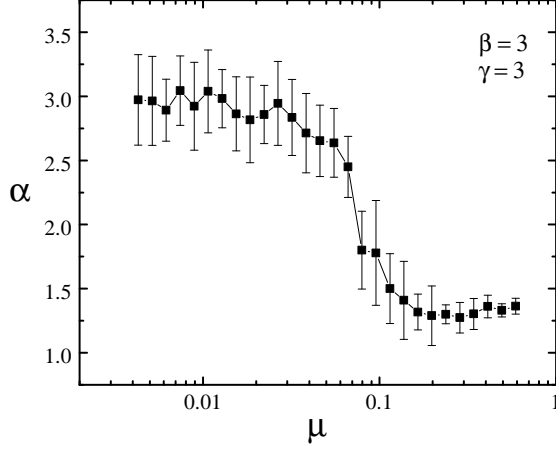


Figure 4: Dependence of the scaling exponent α in the relation $t_{LA_2} \sim N^\alpha$ on the mixing parameter μ . The complexity of the time t_{LA_2} ranges between 1 (overlapping communities) and 3 (clearly separated communities).

$\beta = 2$ and $\gamma = 3$ when $N > 10^4$, and, for $\beta = \gamma = 3$ when $N > 10^3$, i.e. for networks of medium size. Figs. 3 and 4 demonstrate, how the complexity of t_{LA_2} depends on the mixing parameter μ . This is shown in comparison to the other analysed times. In this figure, one can see the remarkable transition from the linear, $\mathcal{O}(N)$, scaling of t_{LA} for $\mu \rightarrow 1$, to the cubic-like, $\mathcal{O}(N^3)$, regime for $\mu \ll 1$.

The above findings suggest possible application areas of the described algorithm. It has a potential to generate large networks under assumption that the communities are fuzzy ($\mu \rightarrow 1$) or that there is a small number of sufficiently large clusters.

Comparing the time execution of the both algorithms one can state that for moderate network sizes the KA outperforms the LA by orders of magnitude. Extrapolating straight lines in Fig. 1 and 2 one can estimate that, for the fuzzy communities, $\mu \rightarrow 1$, t_{KA} becomes larger than t_{LA} for $N > 10^6$, i.e. for really large networks. In the case of well defined communities, $\mu \ll 1$, the time t_{KA} will never exceed the time t_{LA} .

Let us now discuss the quality of the both algorithms. It can be assessed on two levels, namely the level of total node degrees and the level of internal node degrees, see Eq. (8). On the first level, each node, i , of the considered networks is characterized by two parameters: the expected degree, $\langle k_i \rangle$, and

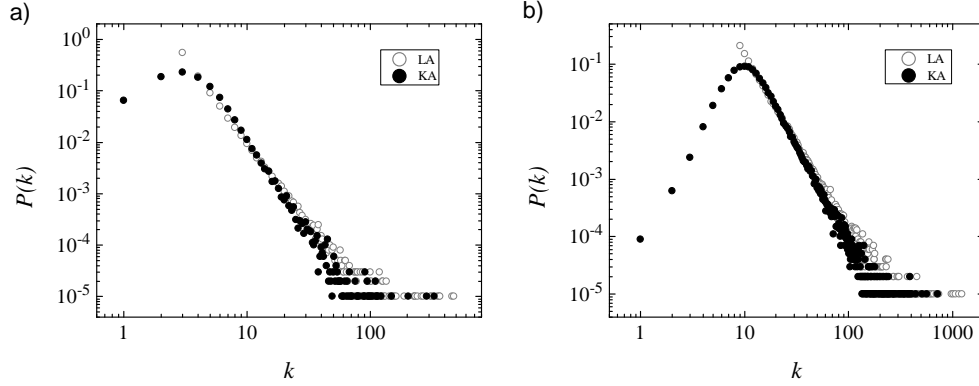


Figure 5: Node degree distributions obtained by numerical simulations using both algorithms, KA and LA. The figure presents results for single networks of size $N = 100000$, $\gamma = 3$, and two different values of the expected node degrees: (a) $\langle k \rangle = 4$, and (b) $\langle k \rangle = 16$.

the obtained degree, k_i . In the case of LA, both these quantities are equal since, after assigning the expected node degrees, one just matches together half-edges emanating from each node. In the case of KA, the probabilistic character of connections between different pairs of nodes leads to asymptotic scale-free networks. The resulting node degree distribution is blurry as compared with the expected one (see Fig. 5). This is due to the fact that the obtained distribution is a kind of convolution of the expected scale-free distribution and the Poisson distribution [23, 24]. The Poisson-like blur of each node degree is the most perceptible for low degree nodes. For medium and large degrees (hubs) it is almost imperceptible.

The direct consequence of the mentioned blur is the occurrence in KA networks isolated nodes. The number of such zero-degree nodes $N_{k=0}$ strongly depends on the average node degree, $\langle k \rangle$, and it can be significant in sparse networks. For example, $N_{k=0} \approx 0.1N$ in the KA network shown in Fig. 5a) for which $\langle k \rangle = 4$, and $N_{k=0} \approx 0.00006N$ in the KA network with $\langle k \rangle = 16$, shown in Fig. 5b). We numerically checked that $N_{k=0}$ decreases exponentially with $\langle k \rangle$.

One also has to keep in mind that in both algorithms, the so-called structural correlations which occur for $\gamma < 3$ may lead to discrepancies between k_i and its expected value, $\langle k_i \rangle$. To avoid them, one has to assure that

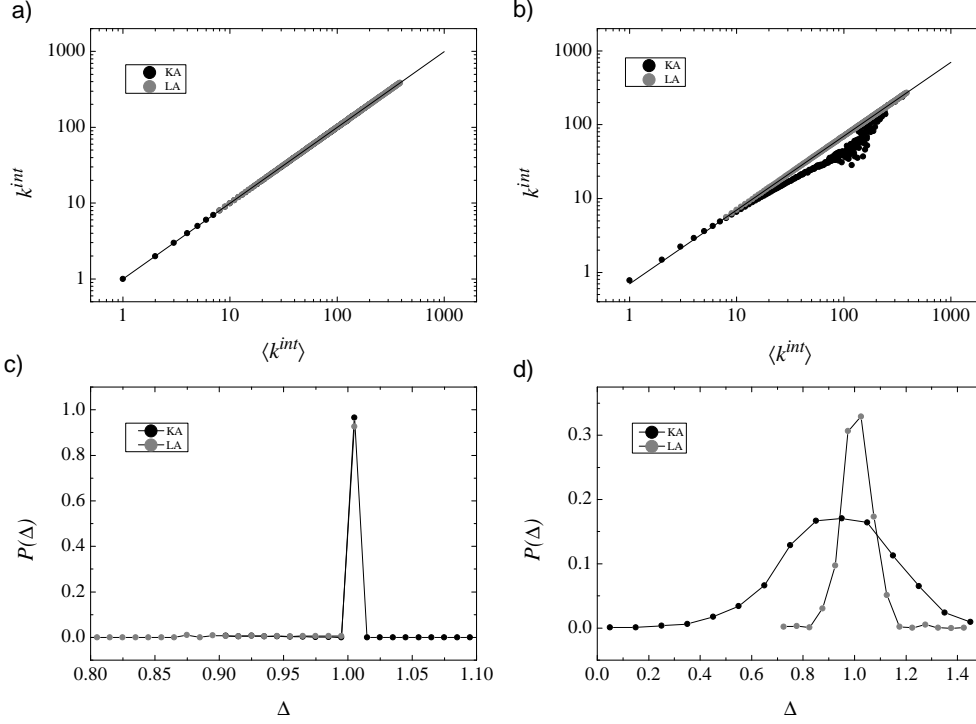


Figure 6: Obtained internal node degrees, k^{int} in relation to their expected values, $\langle k^{int} \rangle$, and the distributions $P(\Delta)$ of the deviation $\Delta = k^{int} / \langle k^{int} \rangle$ for $\mu = 0.01$ (a,c) and for $\mu = 0.3$ (b,d). The results are averaged over 100 realizations of networks with $N = 10000$ and $\gamma = \beta = 3$.

$$k_{max} < \sqrt{\langle k \rangle N} \text{ (cf. Eq. (16) in [21])}.$$

On the second level of the analysis each node of the considered networks can be characterized by the two corresponding parameters: the expected internal degree, $\langle k_v^{int} \rangle$, and the obtained degree, k_v^{int} . The both algorithms can achieve the agreement between the both quantities only in average, however the optimization performed in the step 6 of the LA suggests that this algorithm should be much more precise than the KA in this context. The comparison of both algorithms presented in Fig. 6 confirms this statement.

Deviations between $\langle k_v^{int} \rangle$ and k_v^{int} in KA result from the approximation in Eq. (4), namely from the fact that we neglect the node degree $\langle k_{i,r}^{int} \rangle$ in the denominator of this expression. Such an approximation is crude when the considered node i is a hub (i.e. for $\langle k_{i,r}^{int} \rangle \approx k_{max}$) and the sum of the degrees

characteristics	LA	KA
time efficiency	−	+
quality of the generated networks	+	−
analytical solution	−	+
no zero-degree nodes	+	−
simplicity of implementation	−	+

Table 1: A comparison of the all discussed characteristics of the both algorithms.

230 of the rest of the nodes in the cluster is small. The approximation will work much better if the clusters are dense, i.e. for $\mu \ll 1$. This conclusion is confirmed by the differences in the quality of the KA algorithm for $\mu = 0.01$ (Fig. 6a and 6c) and $\mu = 0.3$ (Fig. 6b and 6d).

5. Conclusions

235 Both algorithms have their own advantages and shortcomings. They are gathered in the Table 1. While choosing an adequate algorithm one has to consider a trade off between accuracy, speed and analytical tractability of the algorithms. It is clear, however, that KA is much faster and allows to generate huge networks ($N > 10^6$) in a reasonable time. It can be easily described
240 analytically (and probably expanded, e.g. taking into account node-degree correlations and even overlaps of communities). On the other hand, LA is much more precise. The variance between expected and obtained node degrees is strongly reduced thanks to the implemented optimization stage. Finally, the algorithms provided in the Appendix, as well as the source codes
245 for KA [25] and for LA [26], clearly demonstrate that the former outperforms the later in term of the simplicity of implementation.

Appendix A. Algorithms reproducing graphs with community structures

250 Here we provide listings of the both algorithms for generating networks with community structures. Due to its complexity, we decided to show only overview of the LA method (**Algorithm 2**). Lines 7, 8, and 9 in the **Algorithm 2** are in fact sub-procedures, and we provide the detailed listing of the second of them only (as the **Algorithm 3**). The reason is that all these

sub-procedures are quite complicated and that we discuss only this second one in the paper in a more detailed way. On the contrary, the **Algorithm 1** presents KA method with all the details.

Algorithm 1 KA algorithm reproducing graphs with community structure

Input: minimal degree k_{min} , maximal degree k_{max} , network size N , heterogeneity coefficients γ and β , mixing parameter μ

Output: graph $G(V, E)$ and map of nodes into clusters f

```

1: function BENCHMARK( $k_{min}, k_{max}, N, \gamma, \beta, \mu$ )
2:   Let  $V = \{v_i \mid i = 1, 2, \dots, N\}$  be a sequence of nodes in a graph
3:   Let  $E = \{(u, v) \mid v, w \in V\}$  be a set of edges
4:   for each node  $v \in V$  do
5:     draw an expected node degree  $\langle k_v \rangle$  from a power distribution  $P_\gamma(k)$ 
6:   end for
7:    $\langle k \rangle \leftarrow \sum_{v \in V} \langle k_v \rangle / N$ 
8:   Let  $C = \{c_i \mid i \in \mathbb{N}\}$  be a sequence of cluster sizes
9:   Let  $f : V \rightarrow \mathbb{S}$  be a map of  $V$  into the set of clusters  $\mathbb{S}$ 
10:  Let  $n \leftarrow 0$  be an initial number of clusters
11:  repeat
12:     $n \leftarrow n + 1$ 
13:    draw a cluster capacity  $c_n$  from a power distribution  $P_\beta(c)$ 
14:     $totalcapacity \leftarrow \sum_{i \leq n} c_i$ 
15:    for each node  $(v_j \mid totalcapacity - c_n < j \leq totalcapacity)$  do
16:       $f(j) = n$ 
17:    end for
18:  until  $totalcapacity < N$ 
19:  for each node  $v \in V$  do
20:    for each node  $w \in V$  do
21:      if  $f(v) = f(w)$  then
22:         $p = (1 - \mu) \langle k_v \rangle \langle k_w \rangle / (\langle k \rangle c_{f(v)})$ 
23:      else
24:         $p = \mu \langle k_v \rangle \langle k_w \rangle / (\langle k \rangle N)$ 
25:      end if
26:      if  $random < p$  then
27:         $k_v \leftarrow k_v + 1$ 
28:         $k_w \leftarrow k_w + 1$ 
29:         $E \leftarrow E \cup (v, w)$ 
30:        if  $f(v) = f(w)$  then
31:           $k_v^{int} \leftarrow k_v^{int} + 1$ 
32:           $k_w^{int} \leftarrow k_w^{int} + 1$ 
33:        end if
34:      end if
35:    end for
36:  end for
37:  return  $G(V, E)$  and  $f$ 
38: end function

```

Algorithm 2 LA algorithm reproducing graphs with community structure

Input: minimal degree k_{min} , maximal degree k_{max} , network size N , heterogeneity coefficients γ and β , mixing parameter μ

Output: graph $G(V, E)$ and nodes-to-clusters assignment \mathbb{S}

```

1: function BENCHMARK( $k_{min}, k_{max}, N, \gamma, \beta, \mu$ )
2:   Let  $V = \{v_i \mid i = 1, 2, \dots, N\}$  be a sequence of nodes in a graph
3:   for each node  $v \in V$  do
4:     draw a node degree  $k_v$  from a power distribution  $P_\gamma(k)$ 
5:     assign an expected internal degree  $\langle k_v^{int} \rangle \leftarrow (1 - \mu)k_v$  to a node  $v$ 
6:   end for
7:   build a preliminary network  $G(V, E)$  using the configuration model
8:   assign nodes to clusters
9:   rewire internal and external connections
10:  return  $G(V, E)$  and nodes-to-clusters assignment  $\mathbb{S}$ 
11: end function

```

Algorithm 3 Procedure that assigns nodes to clusters in LA algorithm

```

1: procedure ASSIGNNODESTOCLUSTERS
2:   Let  $C = \{c_i \mid i \in \mathbb{N}\}$  be a sequence of cluster capacities
3:   Let  $\mathbb{S} = \{S_i \mid i \in \mathbb{N}\}$  be a sequence of nodes' sequences
4:   Let  $S_i = \{s_j^{(i)} \mid j \in \mathbb{N}\}$  be a sequence of nodes in cluster  $i$ 
5:   Let  $n \leftarrow 0$  be an initial number of clusters
6:   repeat
7:      $n \leftarrow n + 1$ 
8:     draw a cluster capacity  $c_n$  from a power distribution  $P_\beta(c)$ 
9:     create empty cluster  $S_n \leftarrow \emptyset$ 
10:     $totalcapacity \leftarrow \sum_{i \leq n} c_i$ 
11:  until  $totalcapacity < \bar{N}$ 
12:  Let  $\mathbb{Z} \leftarrow \mathbb{V}$  be a sequence of nodes currently unassigned to clusters
13:   $trial \leftarrow 0$ 
14:  while  $\mathbb{Z} \neq \emptyset$  and  $n > 1$  do
15:     $trial \leftarrow trial + 1$ 
16:    for each node  $v \in \mathbb{Z}$  do
17:      select a random cluster  $S_i$  from  $\mathbb{S}$ 
18:      if  $\langle k_v^{int} \rangle < c_i$  then ▷ if cluster is large enough
19:        if  $|S_i| = c_i$  then ▷ if cluster is full
20:           $\mathbb{Z} \leftarrow \mathbb{Z} \cup \{s_1^{(i)}\}$  ▷ move 1st node from  $S_i$  back into  $\mathbb{Z}$ 
21:           $S_i \leftarrow S_i \setminus \{s_1^{(i)}\}$ 
22:        end if
23:         $S_i \leftarrow S_i \cup \{v\}$  ▷ add node  $v$  to  $S_i$ 
24:         $\mathbb{Z} \leftarrow \mathbb{Z} \setminus \{v\}$ 
25:      end if
26:    end for
27:    if  $trial > 3N$  then
28:       $trial \leftarrow 0$ 
29:      merge the two smallest clusters into one cluster
30:       $n \leftarrow n - 1$ 
31:       $\mathbb{Z} \leftarrow \mathbb{V}$ 
32:       $\forall S_i \in \mathbb{S} \quad S_i \leftarrow \emptyset$ 
33:    end if
34:  end while
35:  for each cluster  $S_i \in \mathbb{S}$  do
36:    if  $\sum_{v \in S_i} \langle k_v^{int} \rangle$  is odd then
37:      change  $\langle k^{int} \rangle$  of randomly selected node by 1
38:    end if
39:  end for
40: end procedure

```

References

References

- [1] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proc. Nat. Acad. Sci. 99 (12) (2002) 7821–7826.
- [2] G. W. Flake, S. Lawrence, C. L. Giles, F. M. Coetzee, Self-organization and identification of web communities, Computer 35 (3) (2002) 66–71.
- [3] P. Holme, M. Huss, H. Jeong, Subnetwork hierarchies of biochemical pathways, Bioinformatics 19 (4) (2003) 532–538.
- [4] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.

- [5] M. E. J. Newman, Communities, modules and large-scale structure in networks, *Nature Physics* 8 (2012) 25–31.
- 270 [6] B. Karrer, M. E. J. Newman, Stochastic blockmodels and community structure in networks, *Phys. Rev. E* 83 (2011) 016107.
- [7] G. Palla, I. Dernyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature* 435 (2005) 814–818.
- 275 [8] P. W. Holland, K. B. Laskey, S. Leinhardt, Stochastic blockmodels: first steps, *Soc. Networks* 5 (1983) 109–137.
- [9] K. Faust, S. Wasserman, Blockmodels: Interpretation and evaluation, *Soc. Networks* 14 (1992) 5.
- [10] C. J. Anderson, S. Wasserman, K. Faust, Building stochastic blockmodels, *Soc. Networks* 14 (1992) 137.
- 280 [11] T. A. Snijders, K. Nowicki, Estimation and prediction for stochastic block-structures for graphs with latent block structure, *J. Classification* 14 (1997) 75–100.
- [12] E. M. Airoldi, D. M. Blei, S. E. Fienberg, X. P. Xing, Mixed-membership stochastic blockmodels, *J. Mach. Learn. Res.* 9 (2008) 1981–2014.
- 285 [13] A. Goldenberg, A. X. Zheng, S. E. Feinberg, E. M. Airoldi, A survey of statistical network models, *Found. Trends Mach. Learn.* 2 (2009) 1.
- [14] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, *Phys. Rev. E* 78 (2008) 046110.
- 290 [15] P. Fronczak, A. Fronczak, M. Bujok, Exponential random graph models for networks with community structure, *Phys. Rev. E* 88 (2013) 032810.
- [16] J. Park, M. E. J. Newman, Statistical mechanics of networks, *Phys. Rev. E* 70 (2004) 066117.
- [17] A. Fronczak, P. Fronczak, J. A. Holyst, Fluctuation-dissipation relations in complex networks, *Phys. Rev. E* 73 (2006) 016108.

- 295 [18] M. E. J. Newman, *Networks. An Introduction*, Oxford University Press, Oxford, 2010, Ch. 15.2, pp. 565–588.
- [19] C. R. Shalizi, A. Rinaldo, Consistency under sampling of exponential random graph models, arXiv:1111.3054v3 [math.ST] (2011).
- [20] A. Fronczak, Exponential random graph models, arXiv:1210.7828
300 [physics.soc-ph] (2012).
- [21] M. Boguñá, R. Pastor-Satorras, A. Vespignani, Cut-offs and finite size effects in scale-free networks, *Eur. Phys. J. B.* 38 (2) (2004) 205–209.
- [22] M. Molloy, B. Reed, A critical point for random graphs with a given degree sequence, *Random Structures and Algorithms* 6 (1995) 161–179.
- 305 [23] M. Boguñá, R. Pastor-Satorras, Class of correlated random networks with hidden variables, *Phys. Rev. E* 68 (2003) 036112.
- [24] A. Fronczak, P. Fronczak, Networks with given two-point correlations: Hidden correlations from degree correlations, *Phys. Rev. E* 74 (2006) 026121.
- 310 [25] M. Kowalczyk, P. Fronczak, A. Fronczak, A software package to generate graphs by the described algorithm can be downloaded from <http://if.pw.edu.pl/~agatka/benchmark.zip>.
- [26] A. Lancichinetti, S. Fortunato, F. Radicchi, A software package to generate graphs by the lancichinetti algorithm can be downloaded from
315 <http://santo.fortunato.googlepages.com/benchmark.tgz>.